Case Study 3: Determining Significance of Palindromes in CMV DNA

By:

Himanshu Makhija (A09845605) -- Math/Computer Science, B.S.

Hanna Goldman (A11436767) -- Math/Computer Science, B.S.

Shagun Gupta (A91068956) -- Bioinformatics, B.S.

Amey Paranjape (A53218045) -- Electrical & Computer Eng., M.S.

(UCSD, Math 189/289C - Winter 2017)

Table of Contents

Introduction

a) Background

CMV, or the human cytomegalovirus, can be life-threatening for those with suppressed and/or weaker immune systems, and in order to stop the virus scientists are searching for the virus' point of replication.This point is a special place in the DNA of the virus that contain specific instructions on how the virus should replicate. Since DNA is made up of two possible base pairs, what results is a number of combinations of these pairs in long strings, that can sometimes be read as palindromes. What scientists are searching for then is a specific kind of palindrome, a complementary palindrome, that reads the complement of the forward string backwards (an example would be "GGGCATGCCC").

These palindromes can be special, because points of replications are specifically marked by either a long palindrome, or as seen in the Herpes-Simplex form of the virus, a strong cluster of palindromes. Once palindrome clusters are found, the virus DNA is split around the clusters and individually tested to check if it holds the DNA for replication.

The reason we want to perform statistical analysis to determine how rare a palindrome cluster that may present is because testing the individual sites of replication can be very expensive and is generally a time consuming process.

This also begs the question; what is CMV? CMV is a member of the herpes virus family and carries a major risk for people that suffer from any kind of immunosuppression.

b) Objective

Our objective for this case study is to determine whether the known palindromes present a statistical anomaly in that an unreasonable cluster may prove to be the virus CMV's point of replication in its DNA. These possible sites will be recommended to be tested.

c) Survey Methodology

The sequencing of CMV was completed and published in 1990, and during the following year search algorithms were implemented to check the sequence against any patterns that may occur. Since we are dealing with DNA the size of **229,354** base pairs long, the palindromes less than 10 letters long were effectively ignored, resulting in **296** that were at least 10 letters long.

Once the palindrome sites were collected, various histograms were plotted to compare the intervals and their counts. Interestingly, without respect to the different sizes, two palindrome clusters presented near the 93,000th base pair and another in the 195,000th base

pair of the DNA. Therefore, we have reason to believe that one of these two sites may be the point of replication we are after.

To confirm this theory, another histogram was plotted based off a random distribution of palindrome hits along the base pairs in which no pattern was presented. This reinforces our hypothesis regarding the two clusters.

d) Description of Dataset

The dataset comes in a 1-dimensional matrix populated with 296 entries of integers; each marking the midpoint of a palindrome that is at least 10 letters long. The total number of available points in the DNA is 229,354.

e) Further Research

Cytomegalovirus, or CMV, is a genus of viruses in the order Herpesvirales. It has been found that the Herpesvirus genome carry two distinct classes of DNA replication origins; one for production of viral growth and another to maintain the viral genome during latency.

The lytic cycle, which is one means of viral reproduction, has two identifying elements for its site of replication: a binding site for the DNA-binding protein and an adjacent A+T rich region. Structurally, these sites of replication are composed of direct and inverted repeats wherein these repeating and palindromic elements are often pivotal in stimulating replication.

The paper by Marie J. O. Masse et al. on “Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region” (1992) globally analyzes the human CMV genome which revealed 3 regions that were specifically rich in palindromic repeats between 92,100 - 93,500 base pair region. However, they were found to be upstream of the encoding region of ssDNA (single-stranded). This is as expected because this region is composed of the sequences for ribosomal binding site and the promoter sequences which are known to be rich with repeats due to the presence of microsatellites (tandem repeats) that preserve throughout our lifetime and remain pivotal in fields like forensics (e.g., DNA and fingerprinting).

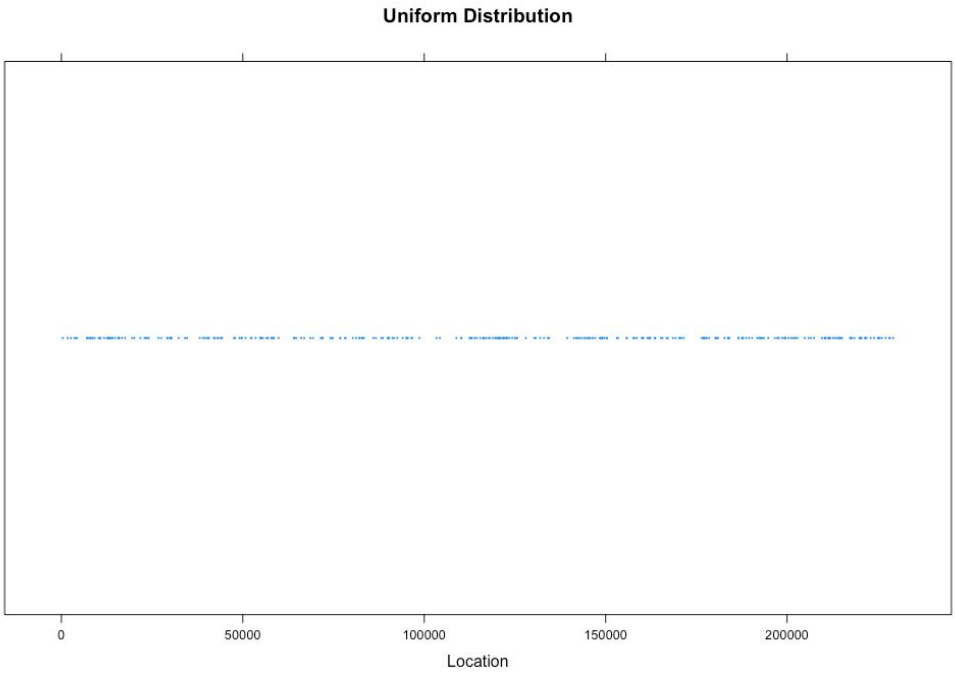
It appears there is increasing evidence of a connection between DNA replication and the expression of adjacent genes. As such, the study by Hermine Mohr et al. (“Cytomegalovirus Replicon-Based Regulation of Gene Expression *In Vitro* and *In Vivo*”) addresses the question of whether a herpesvirus origin of replication can be used to activate or increase the expression of adjacent genes.

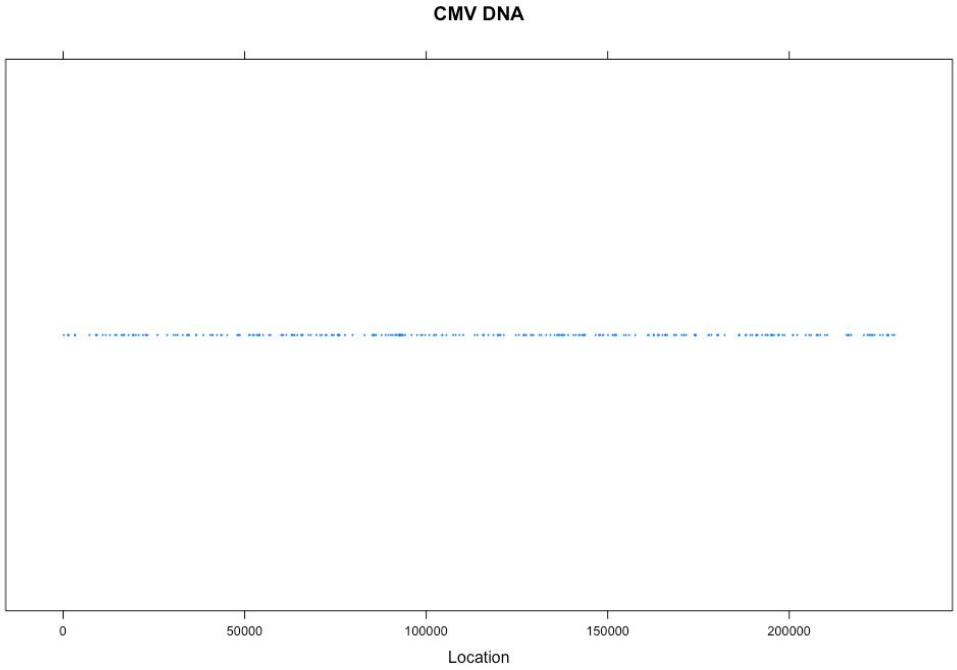
f) Hypothesis

We believe that the palindrome locations along the CMV DNA are based off of mainly uniform distribution with some exceptions for the specific points of replication. These exceptions can be found using analysis procedures on the spacings between each hit (which should follow Poisson process). Therefore, any exceptions to the Poisson process of hits on the DNA should be marked as a possible site of replication.

Analysis

a) Goodness of Fit for Uniform Distribution





As we can see, our CMV palindrome data looks similar to the strip plot generated for the uniform data, where we picked n palindrome sites over N options. To further validate this claim, we perform a chi-square test on our data against the uniform distribution.

To complete the chi-square test, we decided to use Pearson's test:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{k} \frac{(O_i/n - p_i)^2}{p_i}$$

What we have here is a normalized sum of the squared deviations between the theoretical and observed frequencies (Ei and Oi, respectively) and where k is the number of intervals/cells in which the frequencies were calculated. 'n' stands for the total number of samples and finally pi = Ei/npi = Ei/n becomes the expected probability for cell i to get a hit.

First, we calculate observed and theoretical frequencies for our sample and uniform data. Our O values come from the CMV data which is split into 50 regions. For the uniform data, we know that Ei = n/k = 296/50.

Finally, we obtain X^2 = 66.5.

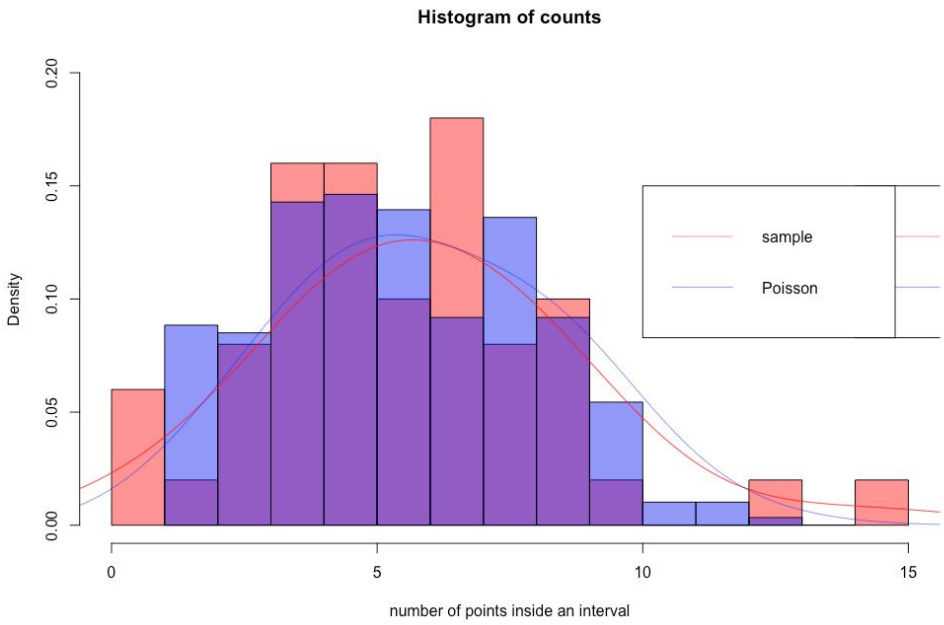
Next, we determine the degrees of freedom, df, of our statistic. In the test of goodness of fit, the degrees of freedom is equal to k-p (where k is number of cells/intervals and p = s+1 is the constraint in the distribution where s is the number of parameters for the distribution we want to test against). In our discrete Uniform case, we have p = 1, and therefore df will be 50 - 1 = 49.

With a desired level of confidence of α = **0.05**, we get a p-value of **0.04864058**. While this value is below the desired confidence value, we _____ (reject/accept?) To check that we were not prematurely rejecting the null hypothesis of our CMV data following uniform distribution, we compare the p-values for different interval numbers to see how the results vary, if they do;

→ In 20 intervals, we have p-value of **0.528**.
→ In 30 intervals, we have p-value of **0.073**.
→ In 50 intervals, we have p-value of **0.048**.
→ In 100 intervals, we have p-value of **0.0034**.

With these results, we can conclude that as we increase the number of intervals, the distribution of our data deviates from the uniform distribution.

b) Goodness of Fit for Poisson Distribution



As we can see, there is a close similarity between the theoretical Poisson distribution and the distribution of palindrome counts in intervals. To check, we performed a chi-squared test to compare the two distributions.

The first step was to divide the CMV DNA into 50 non overlapping regions, each of which had length of about 4587 bases, where we tallied each complementary palindrome into its corresponding interval.

→ 50P (k palindromes in an interval of ~length 4587) = 50e^−λ[1+λ+λ^2/2..+λ^k/k]

Because the rate of hits per interval, λ, is unknown we estimate using the average number of palindromes per interval, which we found to be 5.88 per every 4857 base pairs.
The result from plugging the estimate into the preceding calculations yields **0.112** for the amount of change resulting from 0, 1, or 2 palindromes.

Hence, the approximate expected number of that specific count are as follows:

	Observed	Expected
0	1	0.134260009
1	2	0.794819252
2	1	2.352664987
3	4	4.642592241
4	8	6.871036517
5	8	8.135307236
6	5	8.026836473
7	9	6.788410274
8	4	5.023423603
9	5	3.304296414
10	1	1.956143477
11	0	1.052760853
12	0	0.519362021
13	1	0.236509474
14	0	0.100009720
15	1	0.001231349

For theoretical accuracy, as explained in class, we want each of our rows to have an expected value of at least 5, so we combined rows together, as needed, to get a new table that is more accurate:

	Observed	Expected
0-3	8	7.924336
4	8	6.871037
5	8	8.135307
6	5	8.026836
7	9	6.788410
8	4	5.023424
9-15	8	7.170313

To compare the observed data to the expected, we compute the following:

$$= \frac{(8-7.924336)^2}{7.924336} + \frac{(8-6.871037)^2}{6.871037} + \frac{(8-8.135307)^2}{8.135307} + \ldots + \frac{(8-7.170313)^2}{7.170313} = 2.35487695$$

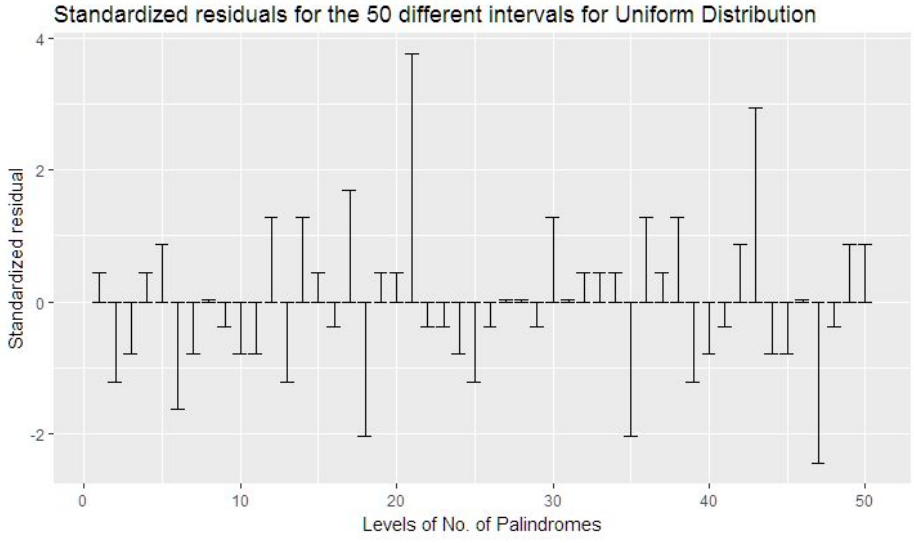
Using the chi-squared distribution to compute the chance of observing a test statistic value at least as large as ours under the assumption of random scatter, we obtain a p-value of **0.79817**. Therefore we cannot reject the null hypothesis (that the Poisson distribution is reasonable).

c) Standard Residual Plot

Residual plots can help us understand and determine where the lack of fit may be occurring. For each category, we plot the residuals using:

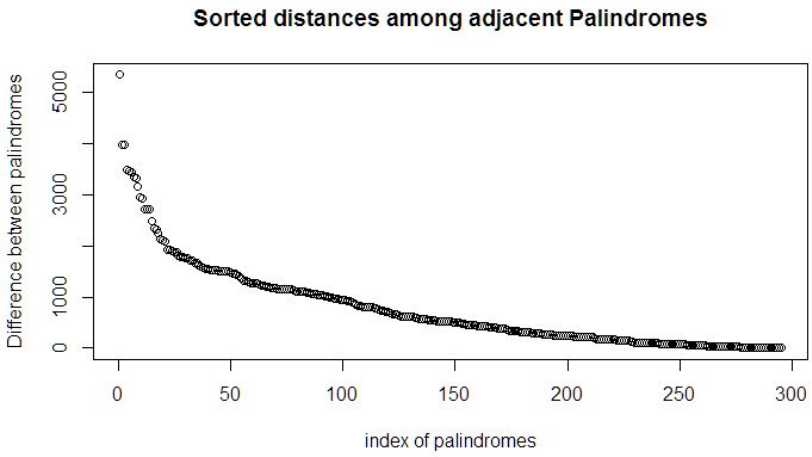
→ $\frac{\text{sample count} - \text{Expected count}}{\sqrt{Expectedcount}} = \frac{N_j - \mu_j}{\sqrt{\mu_j}}.$

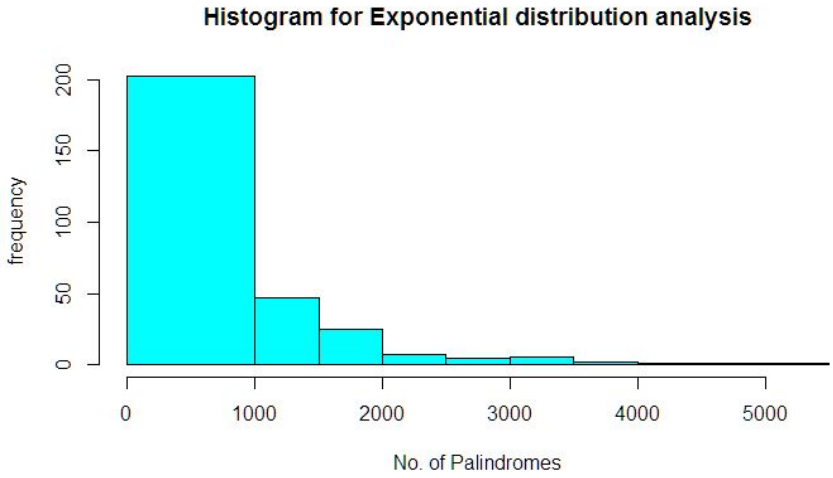
The denominator transforms residuals in order to give them approximately equal variance.



Based on the residual plot above, since anything above the value 3 indicates abnormality, we can say that we observe palindromic clusters that are are relevant (i.e, different from the rest of the dataset, indicating that the data is not a good fit for Poisson at that point) at 21st level of number of palindromes. That is that data is not a good fit in the 21 interval index.

d) Goodness of Fit for Exponential Distribution



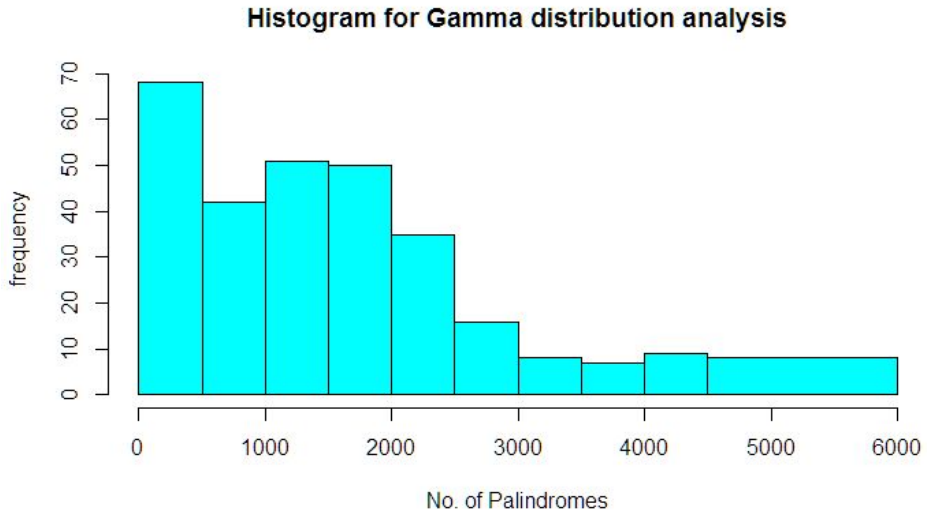


Here we perform the analysis for assumption of Exponential distribution. The value of the chi-squared statistic is **8.153**, with the p-value at **0.319**. We observe the first two bins are the ones contributing the most error, however, qualitatively this graph does seem to follow the exponential curve. The reason this follows is because both Poisson and Exponential variables are related.

Suppose that events (palindromes) occur in space (across DNA) according to a poisson process with parameter lambda. So, random variable X, follows Poisson under lambda and the

probability that the next successive palindrome is exponential with parameter lambda. Hence we expect the graph of differences between palindrome hits to follow an exponential curve.

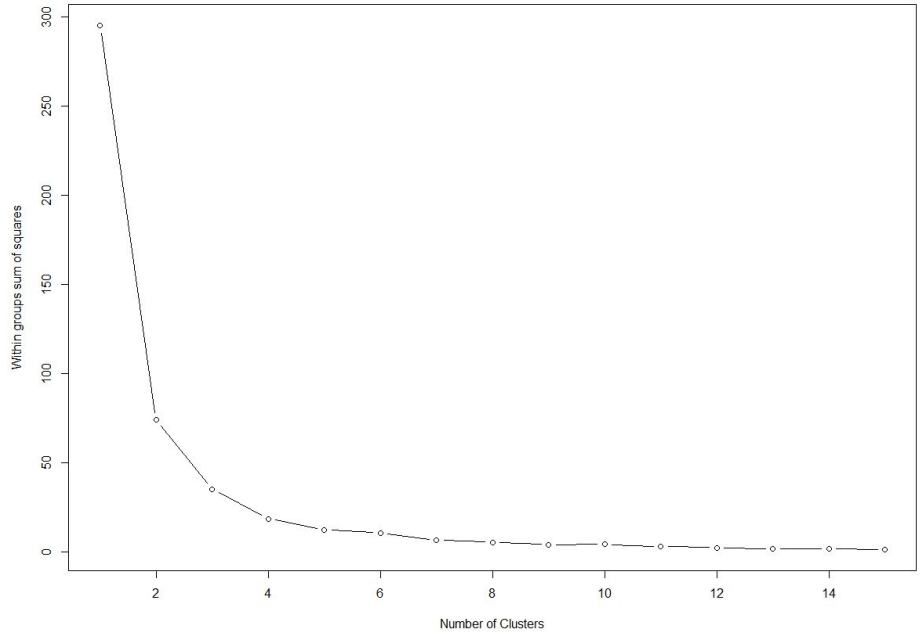
e) Goodness of Fit for Gamma Distribution



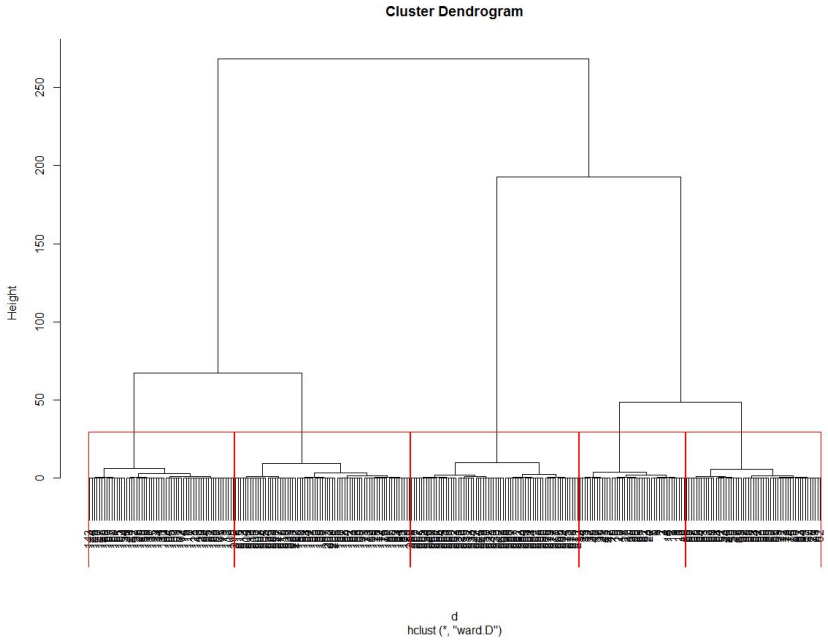
We wanted to check if the distances between palindromic hits are two apart follows a gamma distribution. Here we have a chi-squared statistic value of **infinity**, with a p-value of **0**. Therefore we can reject the null hypothesis that the spacings between two palindromic hits follow a gamma distribution.

f) Advanced Analysis

After preparing the data through listwise deletion of missing values, and then standardizing each entry in the resultant vector of locations, we were able to calculate the number of clusters. Standardization implies a subtraction of the mean of the vector values followed by division by the standard deviation of the same values.



As can be seen from the graph, the curve tells us the most suitable value for the number of clusters, which through the elbow method, we can say the best value in the range [4-5]. This is the point where the curve begins to taper off.

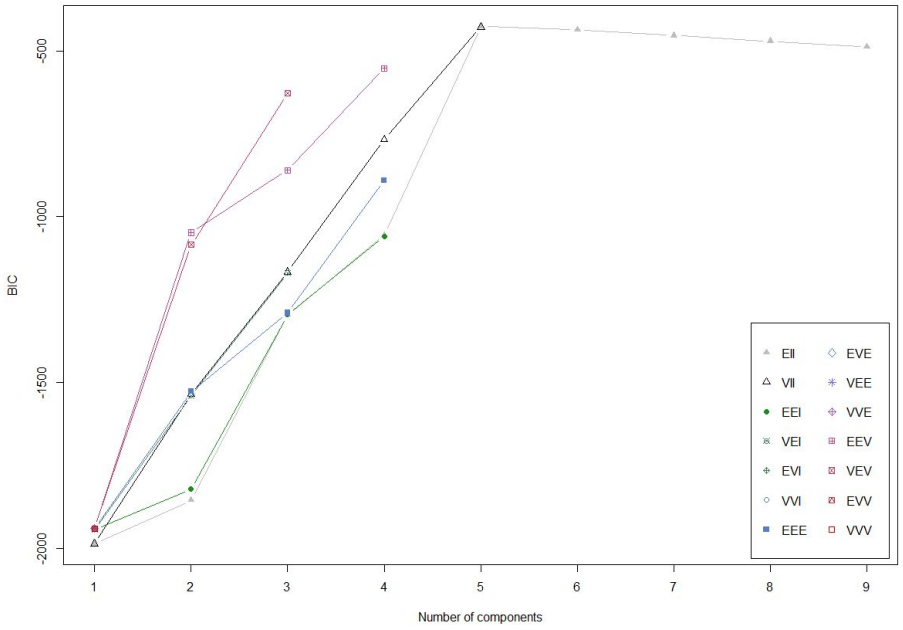


The dendrogram was constructed from hierarchical clustering based on the euclidian distances between the locations of the palindromes. The red lines indicate the distribution of the graph into 5 such clusters. The y-axis is the measure of closeness of either of the individual

points to the clusters. Here we can see that all points in a particular cluster are equidistant to any other point out of the cluster.

Based on prior analysis (the elbow method), we ascertained that 5 was a good measure of the number of clusters for our dataset. This holds in the dendrogram as well, seeing as we have 5 specific clusters allotted.

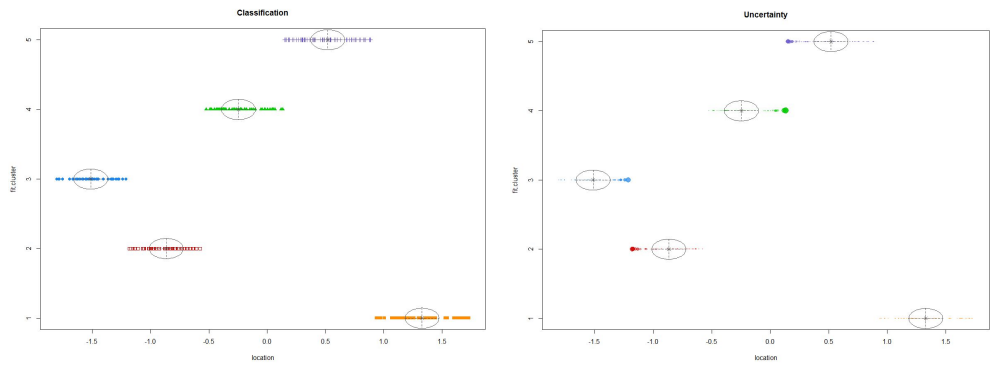
Next, we tested a number of models and used maximum likelihood estimation along with Bayes criteria to identify the most likely model and number of clusters for our dataset.



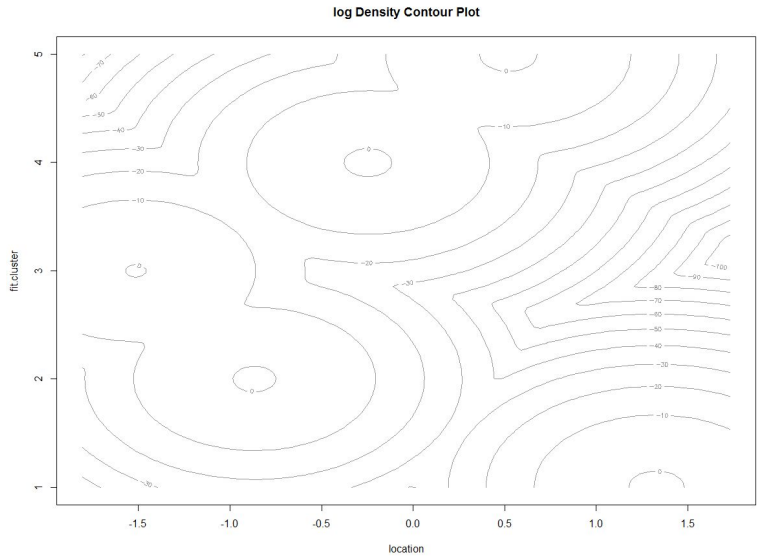
The BIC is a good estimate of what is the best fit model and is similar to likelihood weighting. The lowest value of the BIC indicates the best model, and as such, we have the Gaussian finite mixture model using expectation-maximization algorithm with a BIC value of -426.2014. The number of clusters most suitable for our data is 4 with this method.
Mclust EII (spherical, equal volume) model with 5 components:

```
> log.likelihood   n  df      BIC         ICL
   -170.423     296 15 -426.2014 -426.2014
```

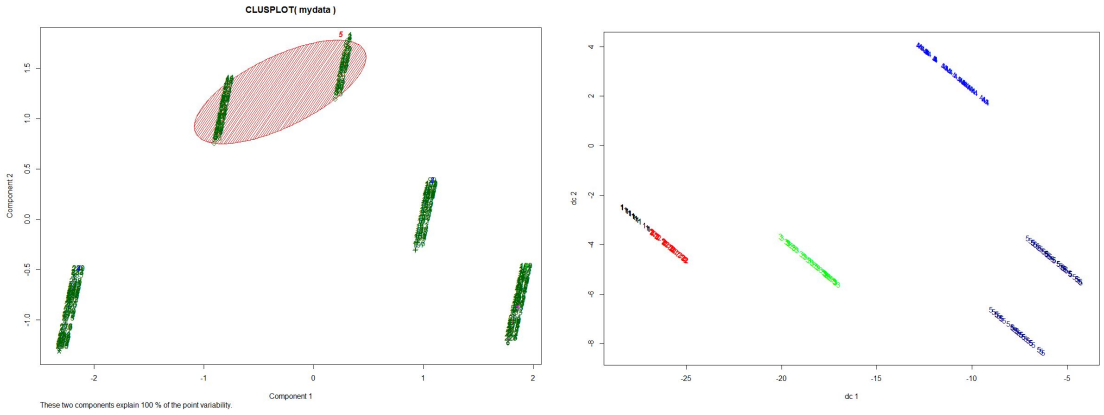
```
>Clustering table:
 1  2  3  4  5
43 55 59 71 68
```



The first graph can help us visualize better the partitions of the dataset based on the chosen clusters, as indicated by the various colors. The second graph is a visualization of the measurement of uncertainty, or deviation, for each of our clusters. It can be said then, that the cluster of data points lying in the range of **1.0** - **1.5** of the standardized location parameter are found more closely versus the other clusters.



Using a graph of the cumulative density function, we again visualize how our data is distributed in terms of density and as can be seen, most of our dataset is found in the **-1.5** - **0.5** region of the standardized location parameter (i.e, 19415-150056)



First, we have a plot after PCA (choosing two principal components from the cleaned and modified dataset). However, both of the graphs above analyze the results obtained after clustering. The higher the percent of point variability of the two components, the greater the information that can be ascertained. Thus because of 100% point variability, the data can be clustered with k-means suitably with the observed two principal components as the means of doing so. From these two, we see the the biggest is in range **-1** - **0.5** (i.e, 109185-150056).

Conclusion

In conclusion, we can say that the data follows a uniform distribution. Based on the results of the standardized residual plot we know that 21st index list, or 21*interval length, which comes close to the 90k region. This is where we get the palindromic sequences that do not fit the Poisson distribution. From the clustering methods, the regions (109185-150056 bp) looks to be the biggest cluster and hence possible site for the origin of replication. Thus, we would recommend the biologist to conduct the experimental trials in the region between 91,000 and 150,000 base pairs for finding the origin of replication and potentially designing a vaccine to treat CMV.

Theory

a) Statistical Theory

We look to a few statistical tools in order to perform the data analysis. First, it is important to distinguish if the distribution of clusters of palindromes is indeed statistically significant. In other words, we must ascertain that clusters occur in a fashion not purely out of chance. We rst have to generate random data to determine which distribution our data best ts. We start by using the uniform distribution and the homogenous Poisson process. The advantages of the homogenous Poisson process include an easy estimate for parameter lambda, a distribution which counts natural random processes, and a parameter which does not depend on the location of the distribution. The expected value of the poisson process is lambda, so lambda captures the underlying rate of number of hits, or underlying rate of number of points along an interval.

Lambda may be approximated using the sample data by calculating the sample average; this estimator is the same using method of moments and maximum likelihood estimation. Note that the probability for k points in an interval is

$$P(k \text{ points in a unit interval}) = (\lambda^k . e^{-\lambda}) / k!$$

For spacing between palindromes, we look to see if an alternate spacing model i.e. (x[i + 2] - x[i]), may be modeled using the Gamma Distribution. The gamma distribution has two positive real numbered parameters that may be parametrized in three di erent manners:

1. With shape parameter k and scale parameter θ.
2. With shape parameter α = k an inverse scale parameter c= 1/θ
3. With shape parameter k and mean parameter μ = k / β

Specifying in terms of k and θ , we have the density function to be

$$(1 / \Gamma(k)\theta^k)\, x^{k-1}\, e(-x/\theta)$$

Notice how the exponential distribution and the chi-squared distribution are special cases of the gamma distribution, with k = 1 and θ = 1, and k = N/2 and θ=2σ2 , respectively. The gamma function is great for modeling random waiting times. Hence, we want to see if spacing between palindromes may be modeled using the gamma distribution.

After simulating data, we must determine which distribution the data best ts. Hence, the chi-squared goodness of t test. Upon partitioning the simulated data into non-overlapping intervals, the chi-squared goodness of t test calculates the standardized deviation of the expected value of that interval based on a distribution to the actual average value of the simulated data in that interval. This calculation is performed for all intervals and summed, then compared to the quantiles of a chi-squared distribution with degrees of freedom equal to,

(Number of intervals) - 1 - (Number of estimators)

The null hypothesis is the data ts a specific distribution versus the alternative that the data does not. For large values of the test statistic, we can expect to reject the null hypothesis as the data deviates from the distribution with a large magnitude.

b) Statistical Methods

1. Homogeneous Poisson Process
    a. A process that can model random phenomena ranging from telephone call arrival times to decay rates of radioactive particles. Mainly arises in processes where points are placed along an axis with no obvious regularities. Some characterizing features of this process follow;
        i. Lambda, which describes the underlying rate of hits per unit on the number line
        ii. Independence of each hit. However, a two hits cannot occupy the same value
    b. In our analysis, the palindromes are simulated using uniform distribution. This means that each interval or base-pair has equal probability of being a hit, thanks to this homogeneous poisson process.
    c. P (k points in a unit interval) = (λ^k)/(k!) * e^(−λ) , for k = 0, 1, · · · …
    d. We expect lambda hits per interval according to our process, but since we cannot know lambda we can use the empirical average number of hits per interval as point estimate for lambda.
2. Chi-Square Goodness of Fit
    a. This test statistic is used to measure how well the distribution fits with the given observed data in mind. It essentially allows you to determine how probable the sample set is to be drawn from the available population. With a high enough probabilistic result, we are able to conclude that the Poisson Process is a reasonable model.
    b. μj = npj , pj = P (an observation is in category j) [n = observed, pj = probability that an observation would be in this category]
    c. We can then compute a p-value, which is the chance that the specific sample set can arise given the probability distribution. If we get a small p-value, we can doubt the probability model used.
3. Residuals
    a. This can show you how a sample set fits into an assumed distribution.
    b. (sample count − Expected count)/(√ Expectedcount)= (Nj − μj)/√μj
    c. Residual values larger than 3 can indicate a lack of fit
4. Exponential Distribution
    a. P (the distance between the first and second hits > t) = P(no hits in an interval of length t) = e^(−λt)
5. Method of Moments
    a. Given an independent sample from a Poisson distribution with parameter lambda, this estimator proceeds as follows:
        i. Find E(X), where X ~ Poisson(lambda)
        ii. Express lambda in terms of E(X)
        iii. Replace E(X) with X-bar to provide estimate of lambda called λ-hat

6. Maximum Likelihood
    a. This method searches among all the Poisson distributions to find the one that places the highest probability to get the observed data using a Likelihood function. The lambda value we want maximizes the likelihood function.
7. K-Means Clustering
    a. We partition the n observations into k clusters. Then we find the best k-value of the clusters for the given dataset.
    b. Given a set of (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, we aim to partition into sets S = {S1, S2, …, Sk} to minimize the within-cluster sum of squares (sum of distance functions of each point in the cluster to the K center).
    c. $$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$
        i. Here ui is the mean of the points in Si.
8. Dendrogram
    a. Built from bottom to top approach, with the idea of grouping the least similar data points together. Eventually, we reach the most dissimilar linkage between all data points in the set.
9. Gaussian Mixture Model
    a. A probabilistic model that assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of latent Gaussians.
    b. EM, expectation-maximization, is an iterative method used in finding the maximum-likelihood of parameters in the dataset.

Works Cited

http://www.pnas.org/content/89/12/5246.full.pdf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3369935/